A Tutorial on

Network Genomics

for the International Conference on

Intelligent Systems for Molecular Biology 2001

by

Christian V. Forst

Bioscience Division,
Mailstop M888,
Los Alamos National Laboratory,
Los Alamos, NM 87545
Tel.: +1 (505) 665-5268
FAX: +1 (505) 665-3024
E-Mail: chris @ lanl.gov

LA-UR-01-2697

1 Introduction

With the ever-increasing genomic information pouring into the databases researchers start to look for pattern in genomes. Key questions are the identification of function. In the past *function* was mainly understood to be assigned to a single gene isolated from other cellular components or mechanisms. Sequence comparison of single genes and their products (proteins) as well as of intergenic space are a consequence of a well established one-gene one-function interpretation. Prediction of function solely by sequence similarity searches are powerful techniques that initiated the advent of bioinformatics and computational biology. Seminal work on sequence alignment by Temple Smith and Michael Waterman [33] and sequence searches with the *BLAST* algorithm by Altschul *et al.*[2] provide essential methods for sequence based determination of function.

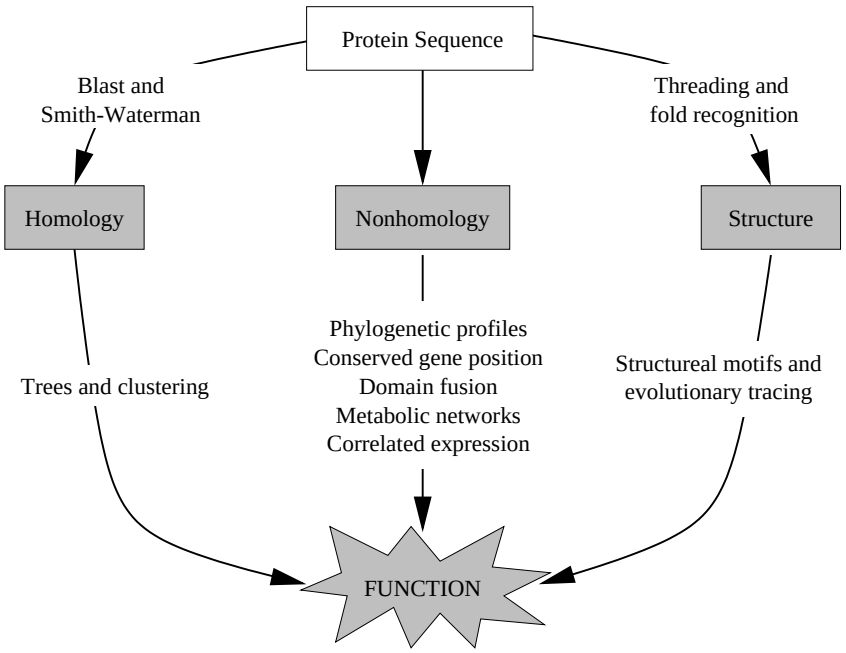

Figure 1: (cf. [21]) Different computational routes to discover the function of a protein.

Similar outstanding contributions to determination of function have been archived in the area of structure prediction, molecular modeling and molecular dynamics. Techniques covering ab initio and homology modeling up to biophysical interpretation of long-run molecular dynamics simulations are mentioned here.

With the ever-increasing number of information of different genetic/genomic origin, new aspect are looked for that deviate from the *single gene at a time* method. Especially with the identification of surprisingly few human genes the emerging perception in the scientific community that the concept of *function* has to be extended to include other sequence based as well as non-sequenced based information.

A schema of determination of function by different concepts is shown in Fig. 1.

The tutorial comprises of following sections; the first two sections will discuss the differences between genomic and non-genomic based context information, section three will cover combined methods. Finally, section four lists web-resources and databases. All presented approaches extensively employ comparative methods.

2 Gene Context

In this section the primary goal will be to discuss different benefits between homology-based and context-based genomic information. The audience will learn the different grades of genome-based contextual information based on genomes such as phyletic profiles, co-occurrence, conserved gene order/conserved operons

and gene fusions. The *Rosetta Stone* approach, based on gene-fusion events will be discussed. I will present techniques how to identify the different grades of genome-based context information and how to use context-information for high-level annotation and analysis. Examples will be used to illustrate the advantages of different context information. The existence of different approach for determination of function is mainly induced by historic events.

In the early days of genomics sequencing a gene or protein was the last step in a tedious and time consuming analysis. Nowadays with fast-track, whole-genome shotgun sequencing the scientific community faces abundant genomic information. Not only DNA and protein sequence information, but also information on internal organization of genes and their location on the chromosome. Such additional information on neighborhood, interaction of functional connectivity is referred to as context information. Based on the

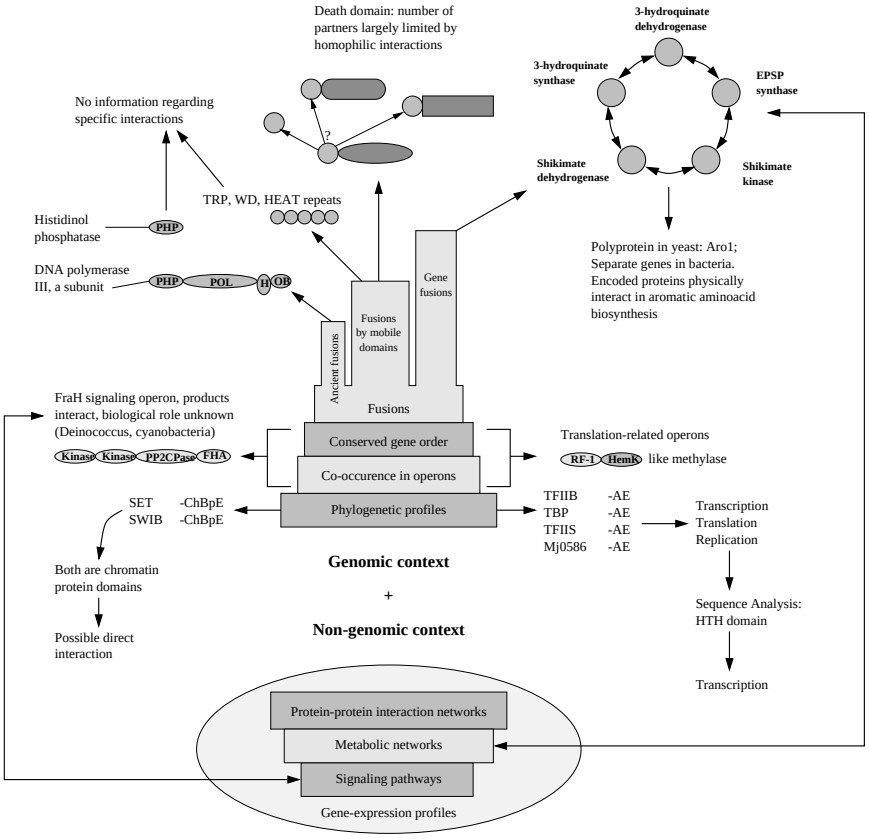

Figure 2: (cf. [4]) A schematic representation of different grades of context information.

source of context information, a rather useful classification is a distinction between genome-based and non-genome based context information. Fig. 2 illustrates different flavors of genome-based context information that relates to non-genomic context information and which will be discussed in the following:

I Phylogenetic profiles/co-occurrence of genes in genomes

II Conservation of local gene neighborhood with two subclasses

IIa Conservation of gene-order

IIb Co-occurrence of genes in operons without conservation of gene-order

III Fusion events

Genomic context information is strongly linked to non-genomic context information, such as protein-protein interaction networks, metabolic networks or signaling pathways (bottom of Fig. 2).

2.1 Phylogenetic profiles/Co-occurrence of genes in genomes

Phylogenetic profiles, the pattern of occurrence of orthologs of a particular gene in a set of genomes under comparison, are the most general form of contextual information [29, 36].

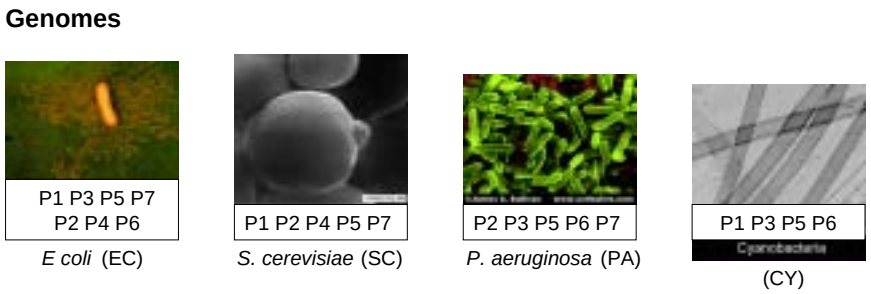

Figure 3: (cf. [6]) Phylogenetic profiles of four hypothetical genomes, each containing a subset of several proteins labeled P1,...,P7. The presence or absence of each protein is indicated by 1 or 0, respectively.

The method of phylogenetic profiles is illustrated in Fig. 3. A phylogenetic profile describes the presence or absence of a particular protein across as set of completely sequenced genomes. If two protein have the same phylogenetic profile, i.e., the same pattern of presence or absence, in all analyzed genomes, it is assumed that the two proteins have a functional link. The example shown in Fig. 3 suggests that P2 and P7 as well as P3 and P6 are functionally linked because protein in both pairs share the same phylogenetic profiles, respectively. Notice that two proteins that are functionally linked in this way, in general, do not share sequence similarity.

2.2 Conserved gene order / Co-occurrence of genes in operons

Identification of conserved neighborhood of genes on a genome provides rather strong evidence of gene interaction. Gradually differences are observed between following gene neighborhoods:

- Gene-clusters; sets of genes where neighboring genes are found in close distance to each other (typically 300bp or less). Genes in gene-clusters do not exhibit particular order or transcriptional orientation
- Conserved gene order; gene clusters that possess a particular order.
- Operons; gene-clusters that do not necessarily exhibit a conserved order of genes, but genes in operons are all oriented in the same transcriptional direction.

Fig. 4 sketches a gene-cluster of two genes (the third gene is not clustered in genome 2 and 3) that is conserved in neighborhood and as an operon. The gene-cluster is not conserved in gene-order assuming all genes are

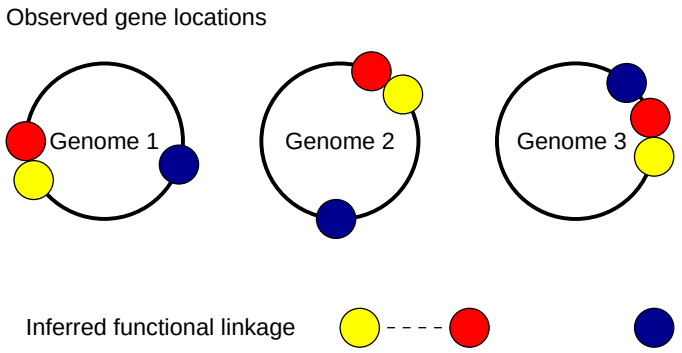

Figure 4: (cf. [6]) Inference of functional linkage by correlating gene neighbors. The light colored genes (red and yellow) are clustered together in all three genomes.

visualized on the plus strand. A more pronounced example in the case of enzymatic genes, functioning in the Tryptophan biosynthesis pathway, is depicted in Fig. 22. A more detailed analysis of this operon as well as pathway will be presented in section 4.2.

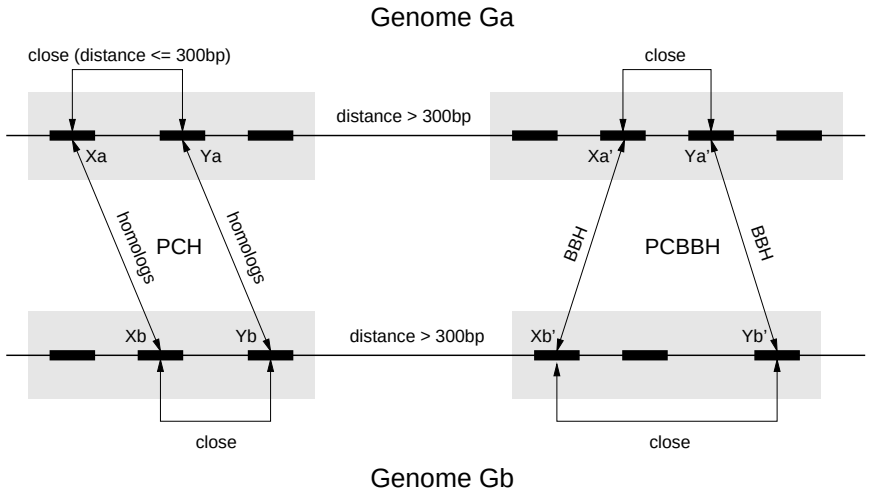

Figure 5: (cf. [28]) Illustration of the definitions of PCBBHs and PCHs (see text).

Overbeek *et al.* [28] introduced a method to identify gene clusters by identifying parallel close hits between neighboring genes of different genomes.

Parallel close hits (PCH) are formed by genes (X_a, Y_a) from genome G_a and (X_b, Y_b) from genome G_b iff X_a and Y_a are close on G_a (closer than 300bp apart), X_b and Y_b are close on G_b, X_a and X_b are recognizable similar and Y_a and Y_b are recognizable similar. "Similar" is defined by Overbeek *et al.* by a FASTA3 score lower than 1.0×10^{-5}. Additional scores are used to value significance of functional coupling provided by PCH. These scores depend on a number of factors, the most important of which is the phylogenetic distance between organisms.

Similar to PCHs, Parallel Close Bidirectional Best Hits are defined: given two genes X'_a and X'_b from two genomes G_a and G_b, X'_a and X'_b are called a *bidirectional best hit* (BBH) iff recognizable similarities exists between these genes (similarity is defined identical as in the case of PCHs), there is no gene Z'_b in G_b that is more similar than X'_b is to X'_a, and there is no gene Z'_a in G_a that is more similar than X'_a is to X'_b. Genes (X'_a, Y'_a) from G_a and (X'_b, Y'_b) from G_b form a *pair of close bidirectional best hits* (PCBBH) iff X'_a and Y'_a

are close, X'_b and Y'_b are close, X'_a and Y'_a are BBH, and X'_b and Y'_b are BBH [28]. Fig .5 illustrates both PCH and PCBBH.

In general, identification and utilization of gene neighborhood is most robust for microbial genomes with their well conserved gene organization. But it may also work to some extent even for human genes where operon-like clusters are observed [41].

2.3 Fusion events

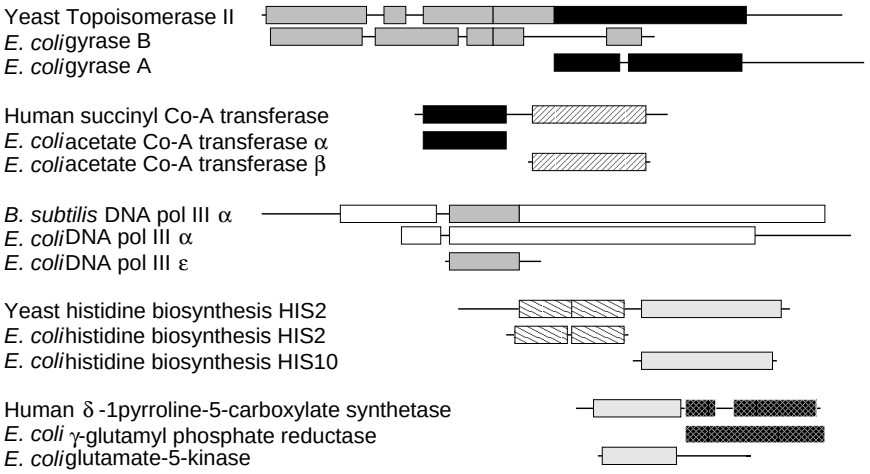

Figure 6: (cf. [20]) Five examples of pairs of *E. coli* proteins predicted to interact by domain fusion analysis.

Fusion genes, the coding of two distinct function on one gene provide powerful information on interacting function. The *Rosetta Stone* approach uses information of fusion events to predict functional linkages between genes. Individual genes in one organisms that are fused into a single chain in another organisms are very likely to interact. Fig. 6 shows examples between non-fused *E. coli* genes and their fusion genes in other organisms.

Gene fusion events are the most effective form of genome context. The encoded proteins of the fused genes tend to be related in function [20], especially if they are orthologs of the fused genes [7, 34].

Marcotte *et al.*[20] provide a hypothesis on the evolution of protein-protein interactions (Fig. 7). Because affinity between proteins A and B is greatly enhanced when A is fused to B, some interacting pairs of proteins may have evolved from primordial proteins that included the interacting domains A and B on the same polypeptide. The shown evolutionary pathway is often referred to as *Rosetta-Stone* hypothesis for evolution of protein interactions.

The domain fusion analysis makes two distinct predictions:

(1) Protein pairs are predicted that possess similar biological functions, e.g., proteins that participate in a common structural complex, metabolic pathway (next chapter) or biological process. Prediction of function is robust; for *E. coli*, general functional similarity was observed in over half the testable predictions [20].

(2) The method predicts potential protein-protein interactions under certain conditions; the method may not find all protein protein interactions (false negative) due to evolution of protein-protein interaction by other methods such as gradual accumulation of mutations. Or it will identify false candidates for interacting pairs (false positives) due to the possible fission (a disappearance of a fusion) after a previously fused protein.

As more genomes are sequenced, there is a higher chance of finding Rosetta Stone sequences.

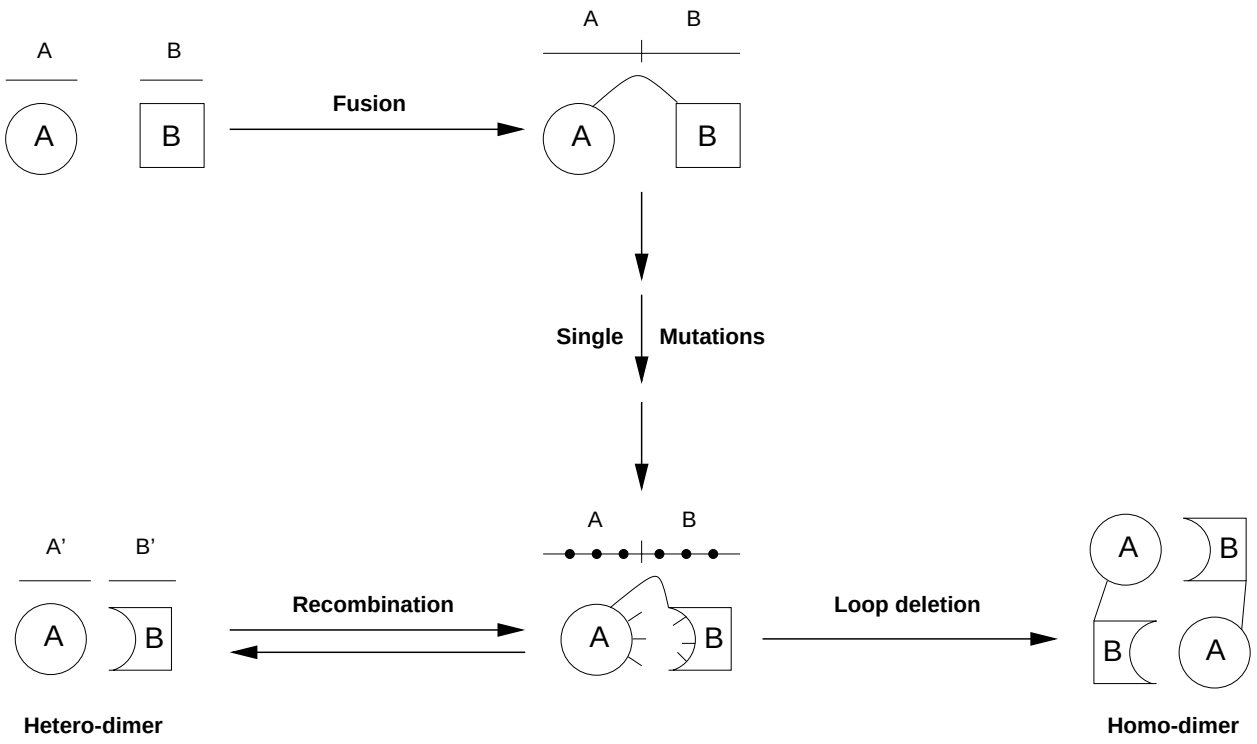

Figure 7: (cf. [20]) A model for the evolution of protein-protein interaction.

2.4 Summary

An interesting statistical survey on different types of functional interaction and their context has been performed by Huynen *et al.*[16]. Functional interactions between proteins of *M. genitalium* have been divided along the previously used hierarchical classification (Section 2).

1. direct physical interaction between proteins
2. indirect physical interaction, i.e., the proteins are part of the same protein complex, but there is no evidence that they interact directly with each other
3. the proteins function in a single metabolic network
4. the proteins function in a non-metabolic network, either regulatory or otherwise
5. the proteins function in the same process
6. pairs of proteins of which at least one is hypothetical
7. proteins with known functions between which no functional interactions are known

A graphical representation of the results is shown in Fig. 9. The surface area of the circles is proportional to the number of genes used in the analysis.

3 Cellular Networks

In contrast to genomic context metabolic and gene-regulatory networks are representatives of non-genomic context information. Protein-protein interaction networks will be included as an intermediate between genomic and non-genomic context information. Different types of networks can easily identified by their

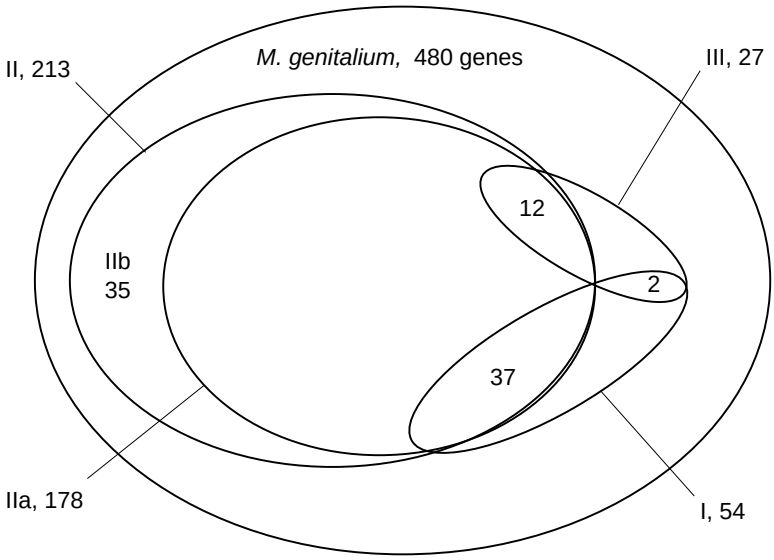

Figure 8: (cf. [16]) Coverage and overlap between various types of genomic context for *M. genitalium* genes. Type I refers to co-occurrence of genes in genomes Type II corresponds to the conservation of local gene neighborhood, which is divided into two subtypes, subtype IIa (conservation of gene-order) and subtype IIb (co-occurrence in operons without conservation of gene-order). Fusion genes are identified as type III.

building blocks, i.e., reactions, as primitive elements. Fig. 10 shows four reactions referring to four types of cellular networks:

- Protein-protein interaction networks
- Metabolic networks
- Signal transduction pathways
- Gene regulatory networks

A successful access to network context is by employing protein function networks and their identification. A powerful approach to identify a subset of protein function networks, i.e., protein-protein interaction is by domain fusion analysis which is addrest in the following section.

3.1 Protein Function networks

The main motivation to identify functionally linked proteins and to induce protein function networks is to predict protein function. Marcotte *et al.*[20] have shown that the general biochemical function of proteins can be inferred by associating proteins on the basis of properties other than the similarity between their amino-acid sequences. These properties associate proteins that are functionally related, i.e., that participate in a common structural complex, metabolic pathway, biological process or closely related physiological function.

By applying methods for the detection of functional linkage to all proteins of an organism, functional network of functionally linked proteins can be mapped out. Fig. 11a shows a network of protein interactions and predicted functional links involving silencing information regulator (SIR) proteins of yeast. Methods that were employed in the construction of this network are experimentally determined interactions, as summarized in the Database of Interacting Proteins [42], interactions predicted by the Rosetta Stone Method (section 2.3) and phylogenetic profiles (section 2.1). Fig. 11b depicts a network of functional links involving the yeast prion protein Sup35 [40].

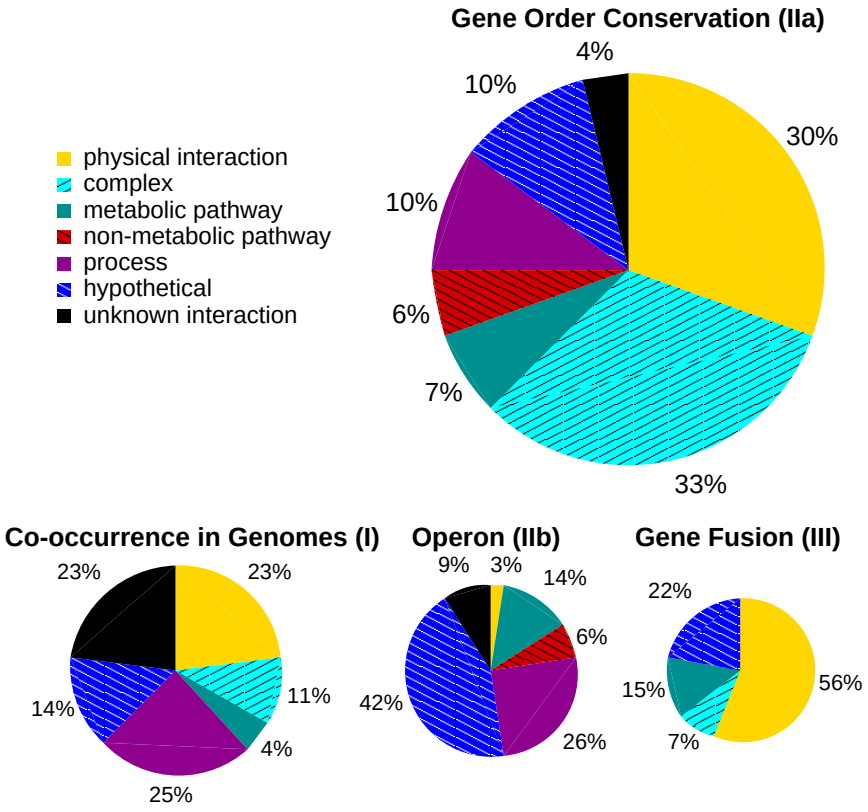

Figure 9: (cf. [16]) Types of functional interaction.

3.2 Metabolic networks

The metabolism of living systems and the evolution of metabolism have been investigated for several decades. The first studies were performed in the late 50s and early 60s by Popper [30, 31] and Lipmann [17]. These studies were followed by others seeking to understand the origin of life and the evolution of the biosphere: seminal contributions by Haldane [13], Miller [24], Oparin [26], and Orgel [27] discussing the (prebiotic) chemical environment suitable for a biotic evolution are noteworthy in this context. Based on these discussions, hypotheses on the origin and evolution of metabolism were formulated [14] and questions regarding the emergence of the first cyclic metabolic networks were addressed, e.g., of the citric acid cycle [38].

In this section we restrict ourselves in reporting the general properties of metabolic networks.

3.3 Gene-regulatory networks

The new and emerging field of gene-expression profiling and gene-expression analysis provides ample and exiting opportunities in computational biology. On the other hand, this research area is in its infancy and, thus, does not provide the researcher with well-established techniques. Although, various techniques are utilized for the analysis of gene-expression experiments.

Gene-expression analysis can be categorized in two types of analysis of different complexity.

- Identification of co-expressed genes
- Inference of gene-regulation networks

Compared to the fast developing gene-expression research area the identification of co-expressed genes is a rather old technique. A spectrum of methods, from simple statistical methods such as calculation of co-

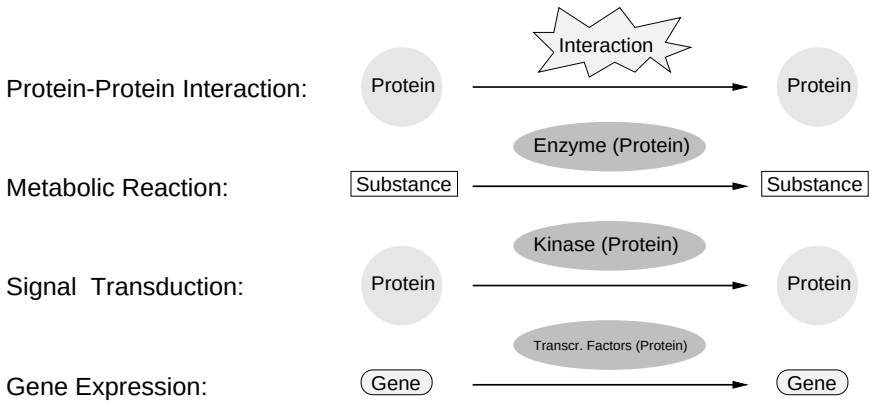

Figure 10: Types of elementary steps in cellular networks.

variance to more complex supervised or unsupervised machine learning technique, e.g., Principle Component Analysis or Support Vector Machines, have been used to classify co-regulated genes

The second type of analysis, the inference of gene-regulation networks uses various techniques covering Bayesian networks, time-series analysis or circuit reconstruction. Main problem in network inference lies in the vast amount of data for statistically significant identification of gene-network connection. The detailed presentation of different gene-regulatory network inference methods exceeds this tutorial. The interested reader is referred to other sources (for example, other tutorials at this conference).

In the lack of sufficient gene-expression data and well established methods we define gene-regulatory network in a broader view based on co-expression. By a combination of techniques to identify functional interactions (methods based on conserved operons, protein fusions and phylogenetic profiles), interacting genes are clustered and the conserved upstream regulatory TIS calculated. Basically a local alignment of clustered upstream sequence motifs have been used and coded in Sequence Logos [32].

3.3.1 Transcription Initiation Sites

A method, not directly related to either gene-content or non-genomic content is the identification of transcription initiation sites (TIS) or *cis*-regulatory elements. Transcription initiation sites are DNA sequences where transcription factors, i.e., proteins, bind. In prokaryotic organisms, TIS are found upstream (*cis*) of the open reading frame (ORF) or coding sequence (CDS) and the promotor site (Fig 12). In eukaryotes TIS for particular genes can be located thousands of basepairs from the corresponding ORF, upstream or downstream. Also TIS were found to be situated in introns of eukaryotic genes.

Different approaches to identify TIS have been pursuit. A method that employs genome context information has been developed in Church's group [22]. McGuire and Church utilize genome based context information, such as conserved operon, protein fusions and phylogenetic profiles. For each of the three methods, matrices of weighted interaction values were calculated, base on the number of genes in the corresponding genome. Greater values indicate predictions of higher confidence. All three interaction matrices are then summed up, and the genes are clustered by the obtained matrix entries in order to obtain predicted regulons. Regulon predictions is performed by local alignment of the upstream sequence regions. Church *et al.* have developed a program, AlignACE [15] for the prediction of TIS sequences from the set of regulons.

McCue *et al.* are pursuing a different approach for identifying TIS. They applied TBLASTN with stringent criteria to identify potential orthologs in genomes of nine gamma proteobacteria. If gene order was conserved, only the sequence intergenic regions was used in the further identification of TIS sequences. Otherwise, the upstream region up to 500bp of the orthologous genes have been used. An advanced Gibbs motif sampler [25] was utilized, that include a motif model of palindromic patters. Also a position specific background model, estimated with a Bayesian segmentation algorithm [18] was used to decide between potential binding site or background.

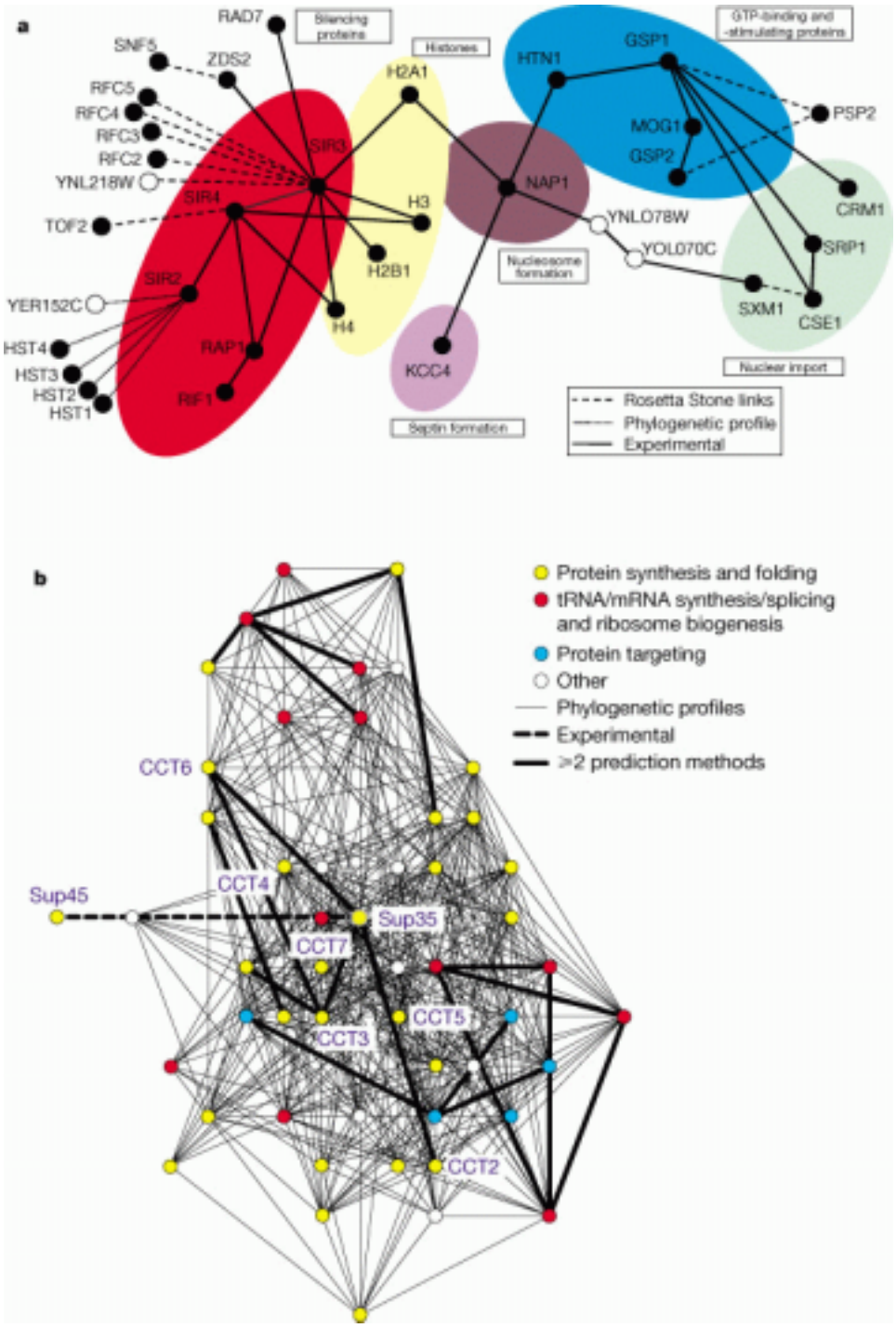

Figure 11: (Reprinted by permission from Nature (Eisenberg *et al.*, *Nature*, **405**, 823–826)[6] copyright (2000) Macmillan Magazines Ltd.) Two functional protein networks of yeast. a) A network of protein interaction involving silencing regulator (SIR) proteins. Filled circles represent proteins of known function; open circles indicate proteins of unknown function. Solid lines show experimentally determined interactions, dashed lines show functional links predicted by the Rosetta Stone approach and dotted lines show functional links predicted by phylogenetic profiles. b) A network of predicted functional linkages involving the yeast prion protein Sup35. The dashed line shows the inly experimentally determined interaction. Solid lines indicated computed linkages (see text). Linkages predicted by more than one method are shown by heavy lines.

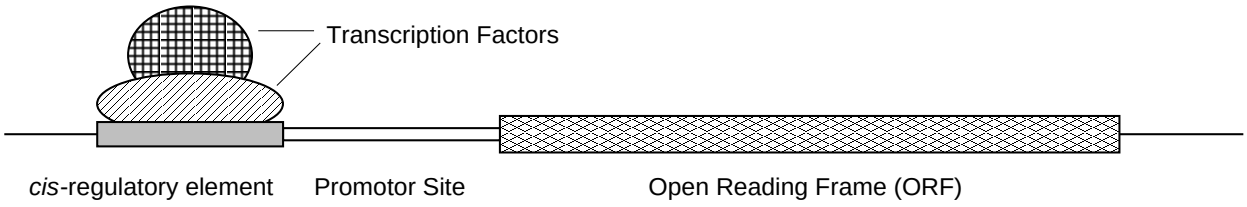

Figure 12: A typical arrangement for a prokaryotic gene. The coding region (ORF) is preceeded by a promotor site and a *cis*-regulatory site.

3.3.2 Principle Component Analysis

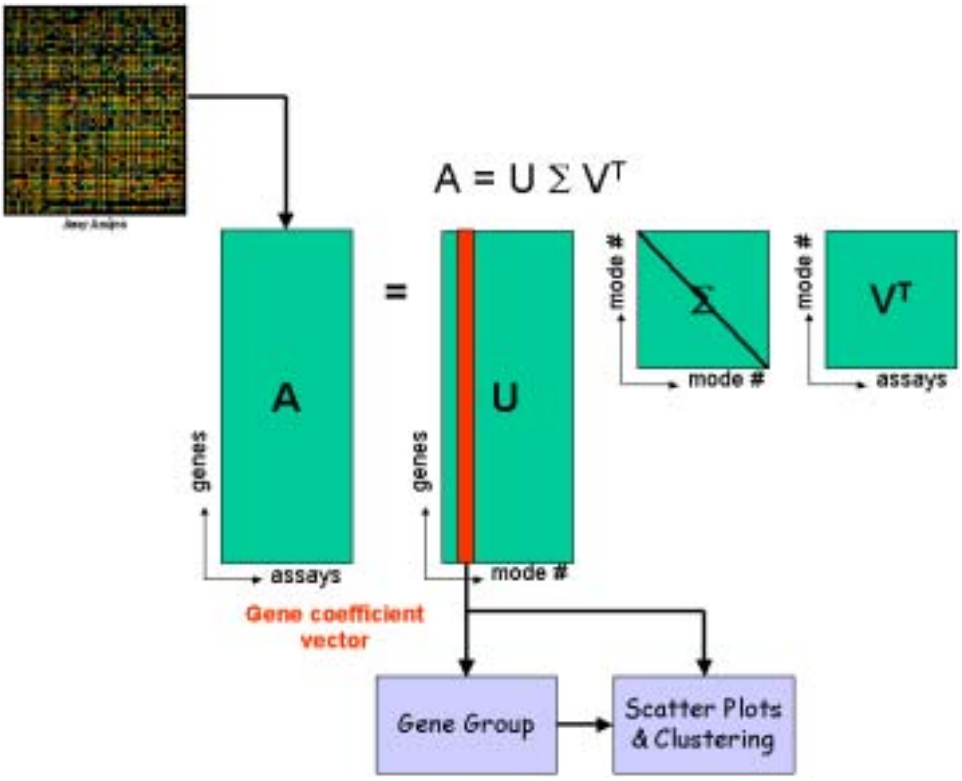

Figure 13: (c.f. [39]) Gene-expression data processing by Principal Component Analysis.

We present a short discourse on gene-expression analysis and use Principle Component Analysis as example [1]. Principle Component Analysis (PCA) [3] is also known as Singular Value Decomposition (SVD) [12] or Karhonen-Loève expansion [19]. PCA is a linear transformation of expression profiles from genes × assays to eigenvectors of genes or gene-vectors × principal components or modes. The gene-vectors are unique up to degeneracy and orthogonal transformations.

The relative expression levels of N genes (for example, all genes of an organism's genome) are simultaneously measured by a single micro-array. A series of M experiments (assays) under slightly different experimental conditions or time-points is then performed. Let the $N \times M$ matrix A donate the full expression data. The PCA is then a linear transformation of the expression data A from the $N \times M$ space into

the reduced $L \times L$ space of singular vectors to modes, where $L = \min(M, N)$

$$A = U\Sigma V^T. \tag{1}$$

The matrix Σ represents a nonnegative diagonal matrix in the reduced space with singular values σ_{ll} corresponding to *eigen-expression levels*. The transformation matrices U and V define the Ngenes $\times$ L modes and the L singular vectors $\times$ M assays, respectively. Fig. 13 shows the schematics of the linear algebra of SVD.

The essential feature of the SVD procedure is to compute the abstract factors so that the factor corresponding to the largest eigenvalue accounts for a maximum of the variation in the data.

4 Combination of Context Information

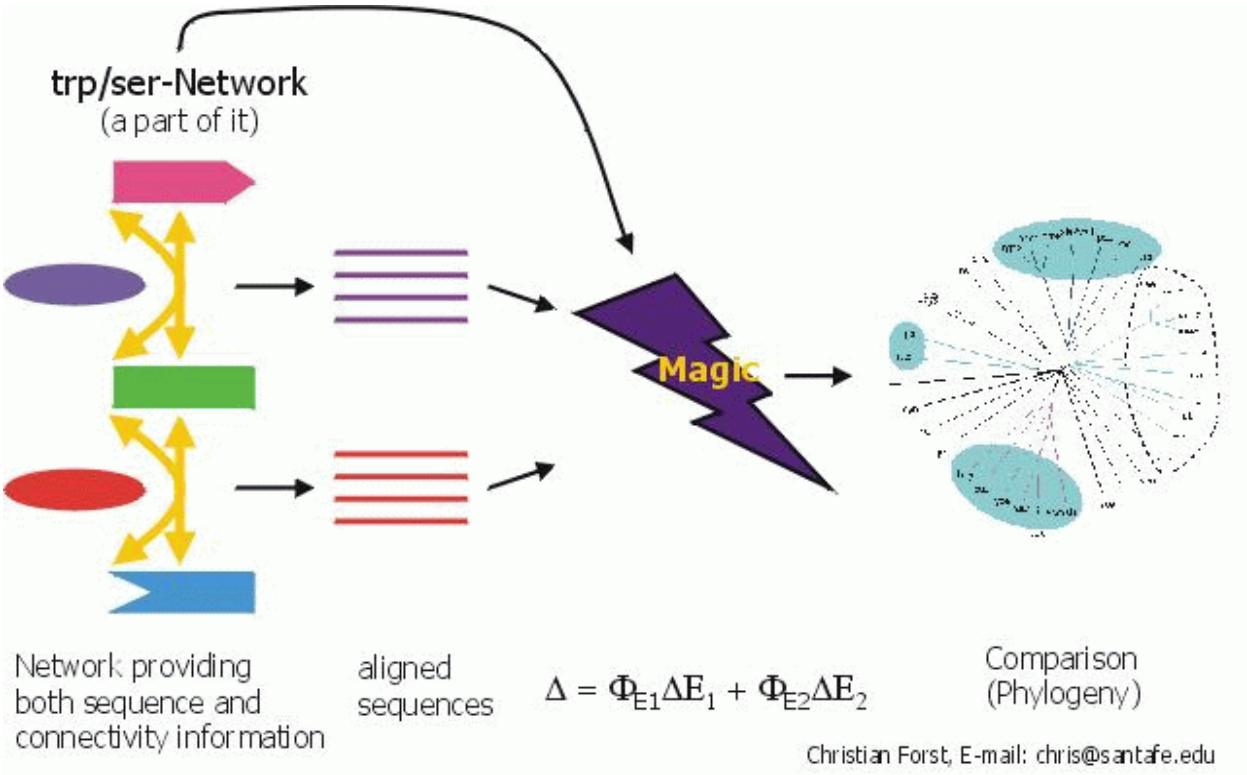

Figure 14: Metabolic network distance. A distance between metabolic networks is defined by using both sequence similarity information from multiple sequence alignment and connectivity information from the metabolic network.

A higher predictive value assemble combined approaches between different grades of context information. Of special interest are combinations between genomic and non-genomic context information for high-level annotation and analysis. Similar to the previous sections, examples are extensively used to illustrate the benefit for bioinformatics research of such combined techniques.

I will make the transition from networks to include genomic based context by referring to the connection between protein-protein interaction and metabolic networks.

With the new technique of comparative network genomics, the quantitative combination between genomic information and network connectivity, relationships between operon conservation and metabolic networks will be discussed. Also, reference to gene-expression pattern of adjacent genes will be made.

Superposition of gene-expression information onto metabolic networks represent the combination of two non-genomic based context information. Main focus for such an approach is the analysis of organismic response to environmental stress.

4.1 Comparative Network Genomics

The analysis of physicochemical properties of metabolic networks is a well established research area that derived from the field of *Origin of Life*. With the advent of post-genomic research and the exponentially growing number of completely sequenced genomes suggest the use of new multi-level approaches that combine genome information with non-genomic network connection. We extend conventional sequence comparison and phylogenetic analysis of individual sequences to metabolic networks. For this purpose we have developed a method that combines distance information of aligned sequences with network information of metabolic networks (Fig. 14) [10, 11]. Connectivity information of metabolic networks is coded into an adjacency matrix and combined with alignments of corresponding enzymes that function in the network. The distance matrices of individual enzymes are then combined by a direct sum, considering gap-distances for missing connections in the network and different weights for ortholog and paralog network presentations.

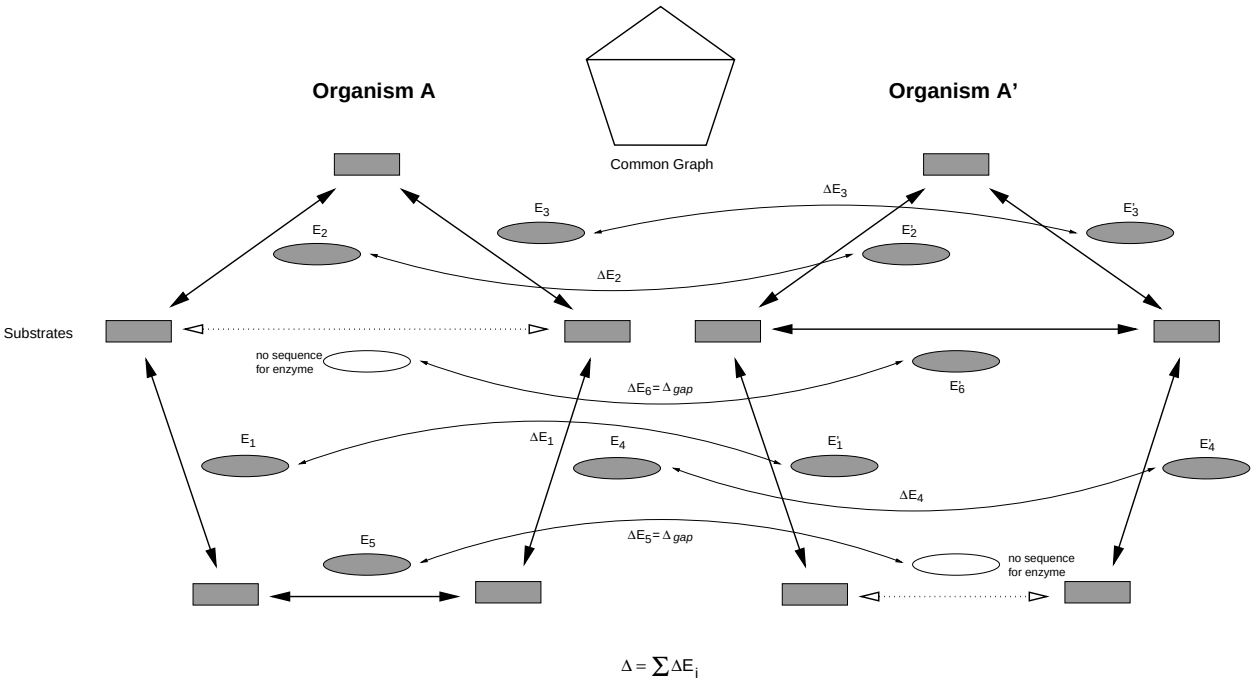

Figure 15: (c.f. [11]) Two networks and their common network. In this example two enzymes E'_5 and E'_6 are not present in both networks. From this results differences in graph-topology between networks of organism A and organism A$'$, one being a cyclic reaction scheme for A that becomes a linear scheme due to an absent enzyme E'_5 in A$'$, and another being a shortcut reaction via E'_6 that is not present in A. Gap penalty Δ_{gap} is assigned to the corresponding distances $\Delta\mathrm{E}_5$ and $\Delta\mathrm{E}_6$.

Consider two networks Γ and Γ' involving n enzymes I_i, I'_i, $i = 1 \ldots n$ and by $\Delta X_i = \delta(I_i, I'_i)$ distances between enzymes I_i and I'_i calculated utilizing an alignment δ. A distance Δ between Γ and Γ' is then defined through

$$\Delta = \sum_{i=1}^{n} \Phi_i \cdot \Delta X_i, \quad \Phi_i \; = \; \begin{cases} 1 & \text{for ortholog pair } i \\ f & \text{for paralog pair } i \end{cases}, \tag{2}$$

where $f > 0$. Different graph-topologies of the network are included in the calculation of distance Δ according to Fig. 15. If a functional role I_k is missing in a pathway Γ then the distance ΔX_k in Eq. (2) is not defined

properly. In this case, to the otherwise undefined distance Δ_k a *gap value* Δ_{gap} is assigned

The above method is employed in the following section for the comparative network analysis of the citric acid cycle and the tryptophan biosynthesis networks. We also remark that the method to calculate network distances is easily adaptable to cellular networks other than metabolic networks.

4.2 From protein-protein interactions to metabolic networks

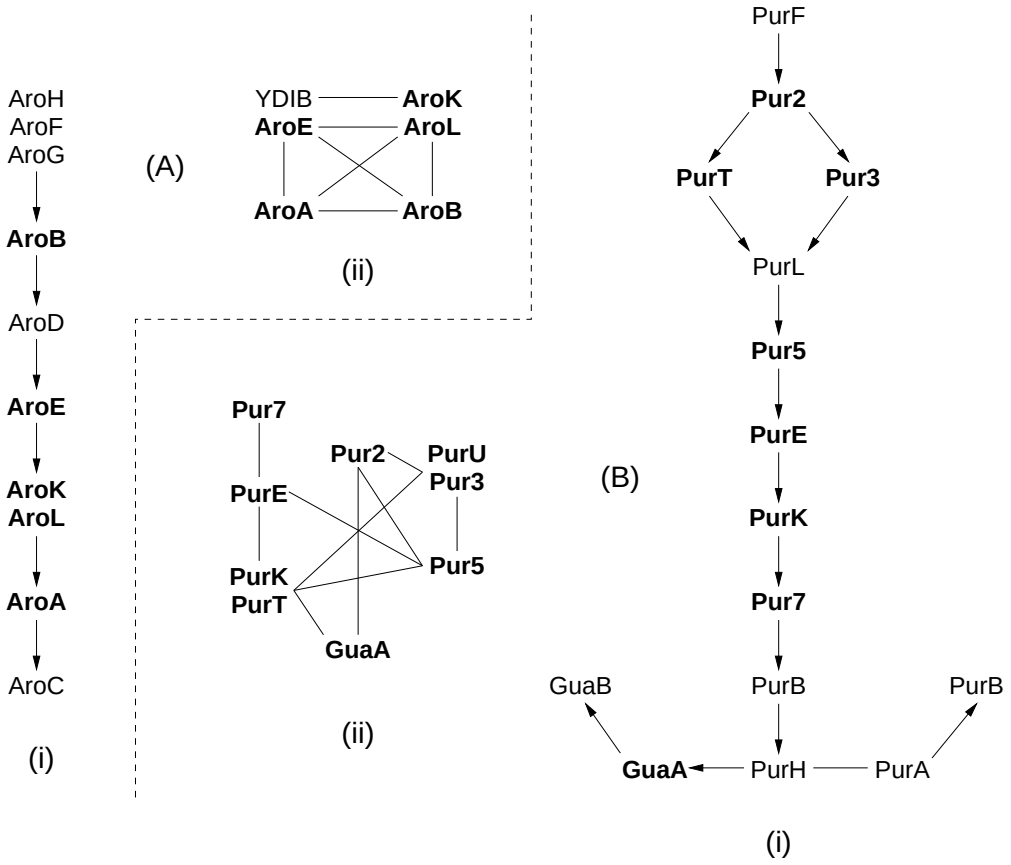

Figure 16: (cf. [20]) Reconstruction of two metabolic pathways in *E. coli* with protein-protein interaction predicted by the Rosetta Stone approach. (i) The pathways studied are known biosynthetic pathways of chorismate (A) and purine (B). (ii) The connection are predicted by the Rosetta Stone method. Enzymes in the pathways that are linked by the Rosetta Stone approach are emphasized.

Previous analysis show that enzymes do interact in protein-protein interaction networks. It is well-known that complex reactions take place in enzyme complexes. Fig. 16 shows two metabolic pathways relevant for biosynthesis of (A) chorismate and (B) purine. Some of the protein-protein interactions are between sequential enzymes in the pathway, and others are between enzymes not directly connected by a reaction, suggesting a multienyzme complex.

4.2.1 Citric Acid Cycle

Another example is the 2-oxoglutarate dehydrogenase complex in the citric acid cycle that converts 2-oxoglutarate into succinyl-CoA. The complex itself consists of 12 2-oxoglutarate-decarboxylase subunits, 24 transsuccinylase units with a lipoamid group each, and 12 dihydrolipoyl-dehydrogenase units. This leads us to a well known example of the Citric Acid Cycle (Krebs Cycle or TriCarboxy Acid, TCA-cycle) where

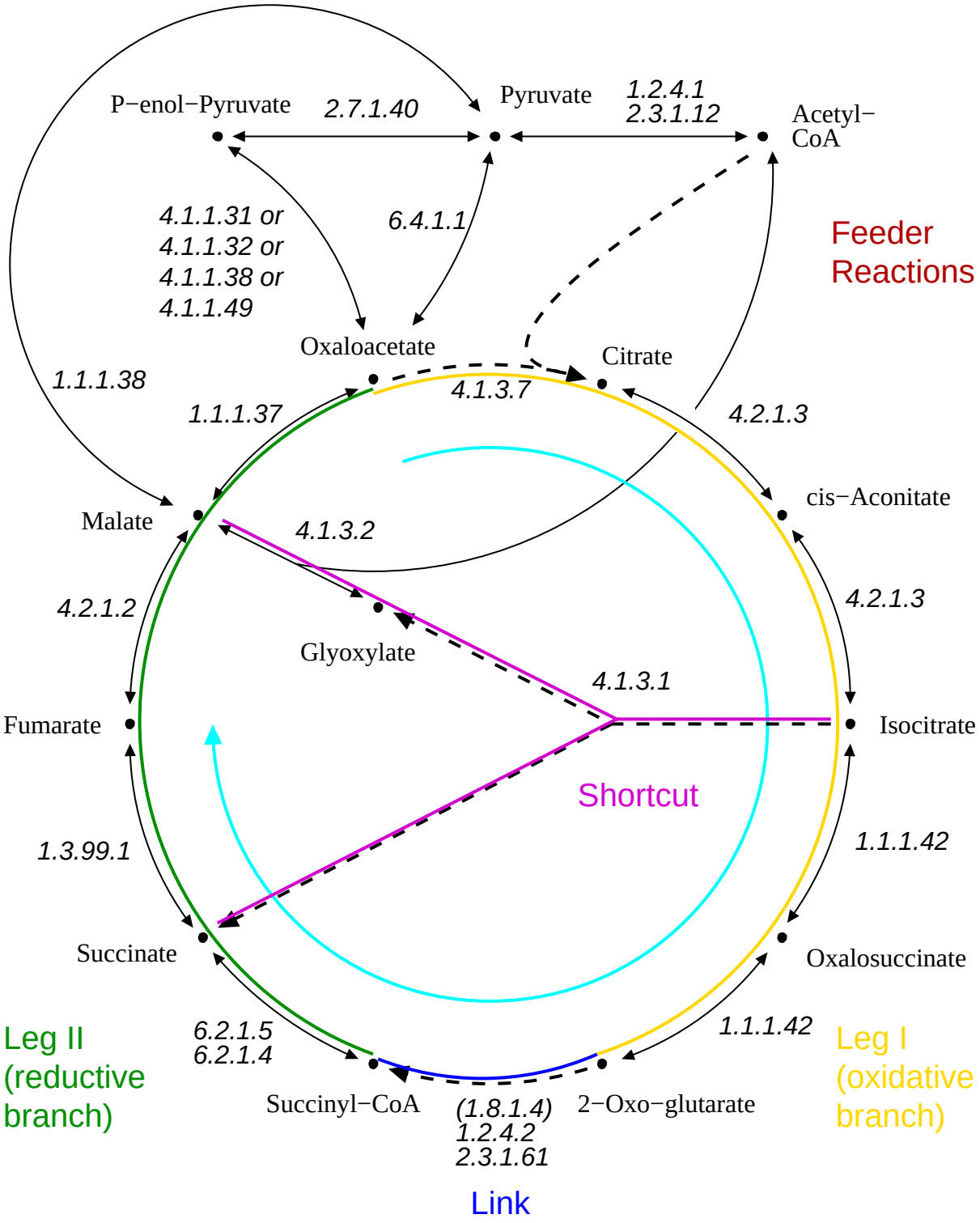

Figure 17: (cf. [10]) The citric acid cycle. The network is divided in feeder reactions from P-enol-pyruvate, pyruvate and acetyl-CoA, in an oxidative branch (Leg I), in an reductive branch if standalone (Leg II), in a link reaction between 2-oxo-glutarate and Succinyl-CoA and in a shortcut reaction via Glyoxylate.

2-oxoglutarate-decarboxylase plays an essential role in the cyclic reaction. The network of the Krebs cycle is

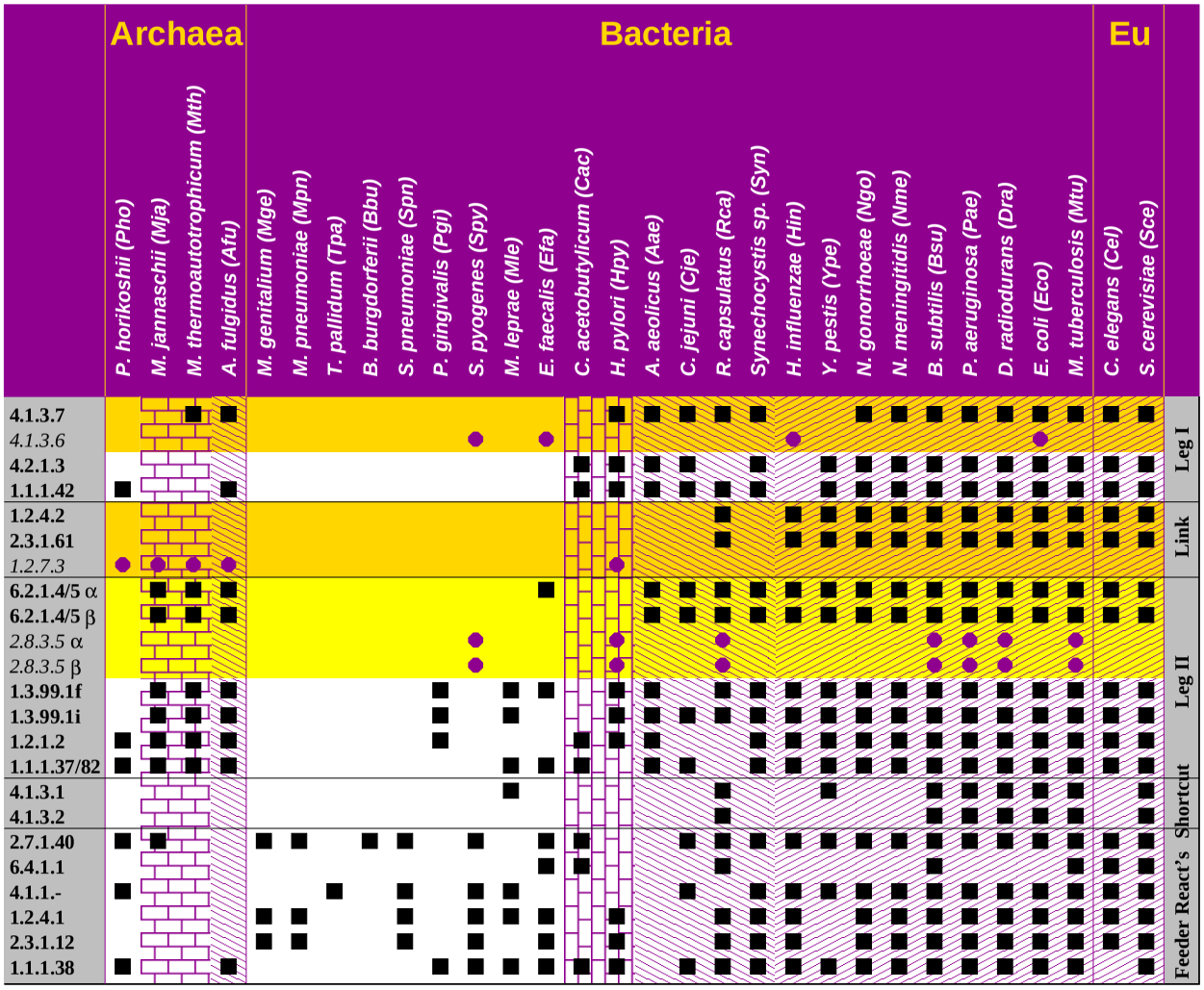

Figure 18: Metabolic profile of the citric acid cycle. Enzymes are group according to the classification of reaction in Fig. 17. Circles refer to enzymes that catalyze reactions in the direction opposite to the usual metabolic flux in the oxidative TCA cycle.

depicted in Fig. 17. A metabolic profile performed for 30 completely sequence genomes is shown in Fig. 18. One can observe the changing metabolic capacity by changing "life-style" between organisms. Autotroph, aerobic organisms do have the complete TCA cycle (bottom left to top right hatching). Some organisms miss the link reaction by 2-oxo-glutarate dehydrogenase (top left to bottom right hatching). The methanogenic archaea (horizontal bricks) only have the reductive branch, *C. acetobutylicum* and *H. pylori* (vertical bricks) only possess the oxidative branch of the cycle. All other organisms have a rather rudimentary metabolic profile of the TCA cycle.

Employing comparative network genomics as outlined in the previous section yields following network similarities visualized as phylogenetic tree (19).

4.2.2 Tryptophan Biosynthesis Network

An interesting relationship between metabolic networks, operon conservation and gene fusion events have been observed in the case of the Tryptophan biosynthesis network (Fig 20). In the last step of tryptophan

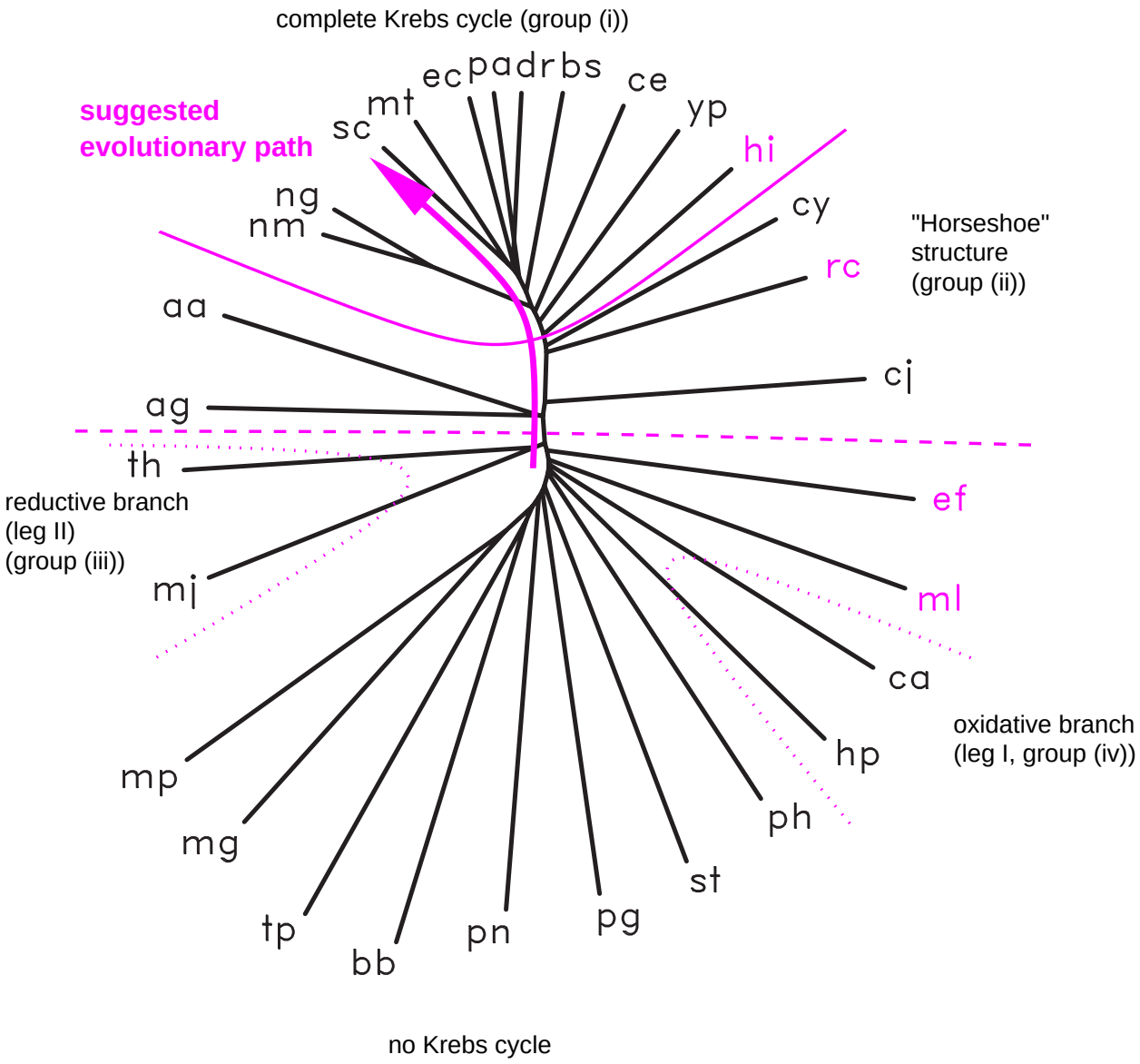

Figure 19: (cf. [10]) Citric acid cycle phylogeny. Network distances are visualized as phylogenetic tree (see text)

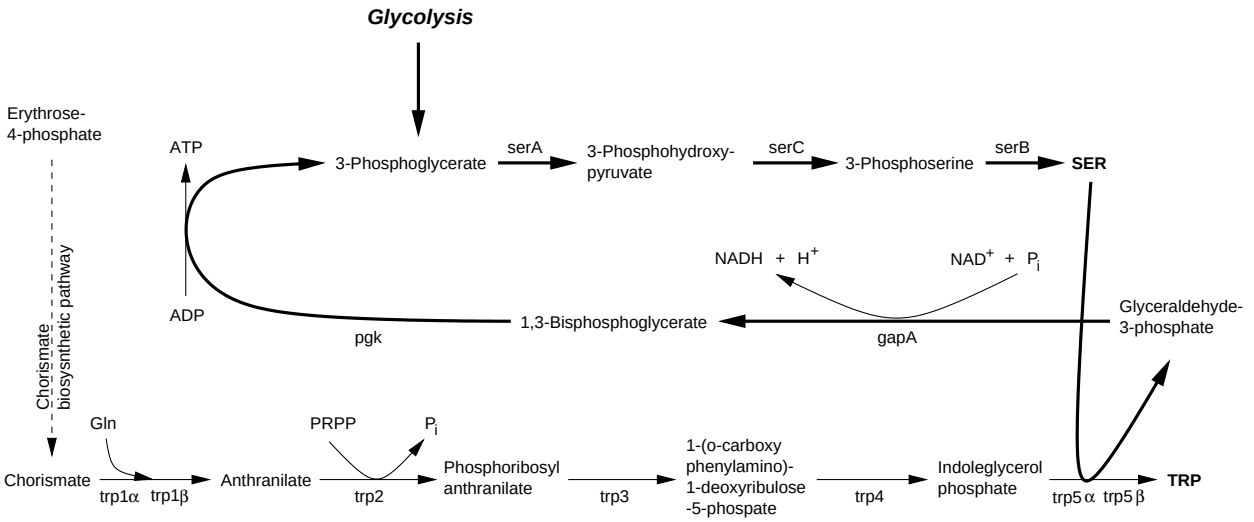

Figure 20: (cf. [43]) Tryptophan biosynthesis network. The interconnected biosynthesis pathway of tryptophan together with the serine biosynthesis pathway and the serine salvage pathway are displayed.

biosynthesis, serine combines with indoleglycerol phospate to produce tryptophan and glyceraldehyde-3-phosphate. The two glycolytic enzymes that are present in almost all organisms, glyceraldehyde-3-phosphate dehydrogenase (gapA) and phosphoglycerate mutase (pgk), recycle the three-carbon glyceraldehyd-3-phosphate to 3-phosphoglycerate. The latter is then transformed via phosphoglycerate dehydrogenase (serA), phosphoserine transaminase (serC) and phosphoserine phosphatase (serB) to serine. Tryptophan itself is synthesized from chorismate via anthranilate synthase component α and β (trpE and trpG), anthranilate phosphoribosyl transferase (trpD), N-(5'-phosphoribosyl)anthranilate isomerase (trpF), indole-3-glycerol phosphate synthase (trpC), and tryptophan synthase α and β chain (trpA and trpB). The metabolic profile of the network is presented in Fig. 21. Here the lack of serC in the archaea is prominent. Although, experimental evidence exists that archaea use the standard phosphorylating pathway to synthesize serine (Stauffer, 1983; Metcalf *et al.*, 1996). Apparently archaea possess serC genes that are unrelated to any serC sequence presently in the sequence databases.

A comparison of the operon organization (Fig 22) with the corresponding network phylogeny (Fig 23) [11] shows the following: Based on the 16S rRNA tree, *E. coli* is closely related to *Y. pestis*, *H. influenzae* and *P. aeruginosa* as shown in Fig. 22. On the other hand, in terms of pathways, *P. aeruginosa* is closely related to *R. capsulatus* as shown in Fig. 23 (clade I) with similar operon organization (*trpE-trpD-trpC...trpF-trpB*) (Fig 22). Closely related *E. coli*, *H. influenzae* and *Y. pestis*, based on the 16S rRNA tree, exhibit very similar pathways. *H. pylori*, distantly related to the former organisms based on the 16S rRNA tree, joins the group in the pathway phylogeny (Fig. 23, clade III) with a common operon organization showing a gene-fusion between trpC and trpF (trpE-trpG-trpD-trpC/F-trpB-trpA) (Fig 22).

Another example for a difference between pathway phylogeny and 16S rRNA tree is observed between archaea and bacteria. *M. thermoautotropicum* (operon: *trpE-trpG-trpC-trpF-trpB-trpA-trpD*) (Fig 22) shows a pathway as well as operon structure that is close to that of *T. maritima* and *C. acetobutylicum* (*trpE-trpG-trpD-trpC-trpF-trpB-trpA*, (Fig 22)) as shown in Fig. 23 (clade II). Only *trpD* changed place during evolution between *M. thermoautotropicum* on the one hand and *C. acetobutylicum*, *T. maritima* on the other hand. At comparison of operon structures of trp-genes for organisms in clade II and of those in clade III suggests a gene-fusion event between trpC and trpF genes. Non-fused trpC and trpF genes in clade II involving *C. acetobutylicum*, *M. thermoautotropicum* and *T. maritima* have been fused during evolution and are exhibited as fusion genes trpC/F in *E. coli*, *H. influenzae*, *H. pylori* and *Y. pestis* in clade III. The gene fusion occurs between the gram-positive bacterium *C. acetobutylicum*, the thermophile bacterium *T. maritima*, the archaeon *M. thermoautotropicum* (clade II) and gram-negative bacteria *E. coli*, *H. influenzae*,

Figure 21: Metabolic profile of the tryptophan biosynthesis network

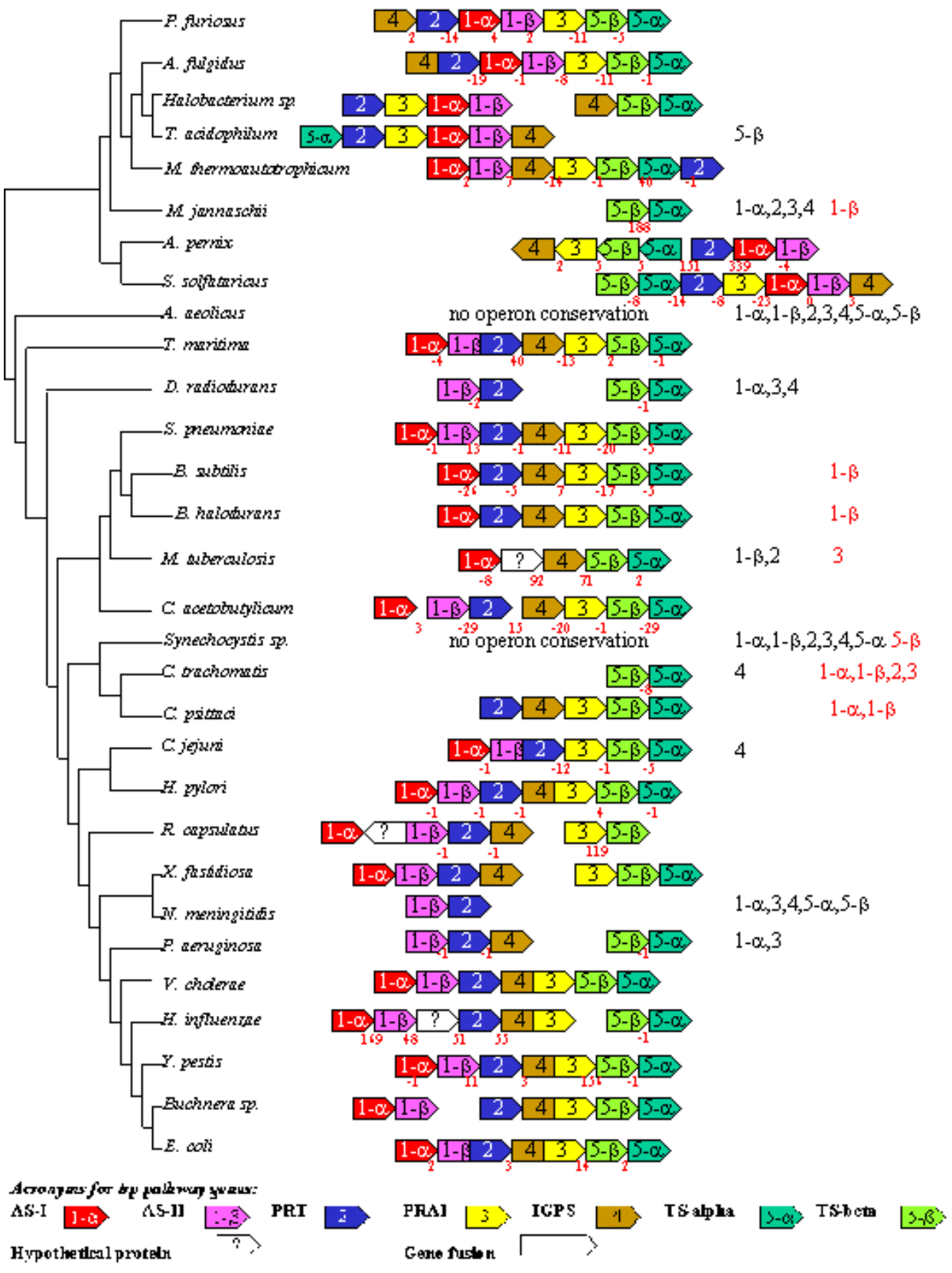

Figure 22: 16S rRNA dendrogram and operon organization. A dendrogram of microbial organisms based on their 16S rRNA together with the operon organization of the *trp*-operon is shown. Genes coding for enzymes in the *trp* biosynthesis pathway are numbered according their step in the pathway, from 1 to 5. For step 1 and 5, respectively, each two subunits α and β are necessary to form a functioning enzyme complex.

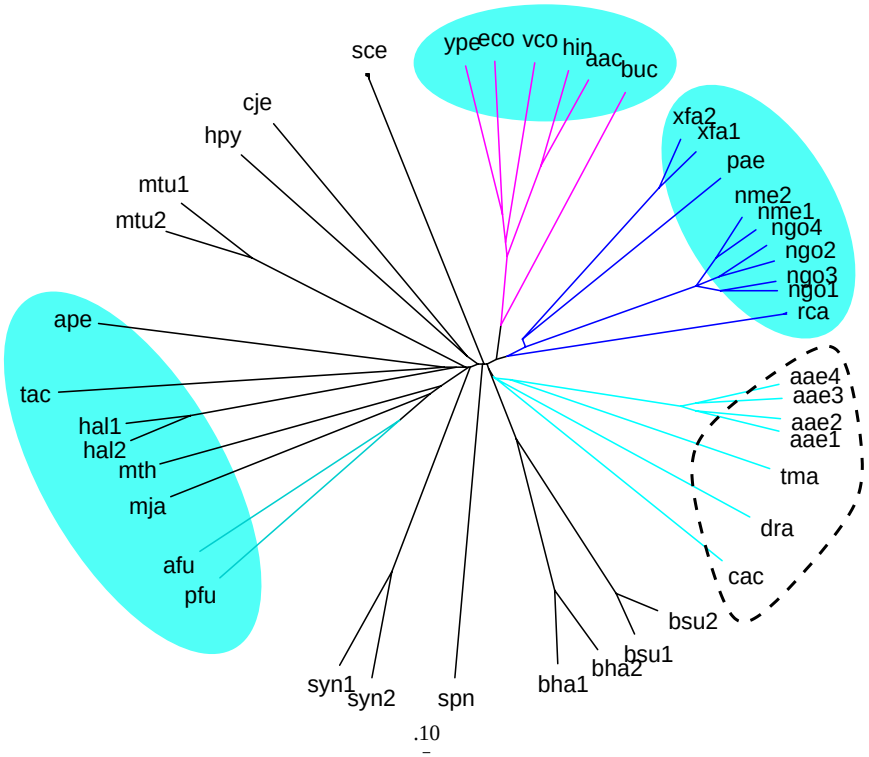

Figure 23: (Tryptophan biosynthesis pathway. The phylogenetic tree is computed with parameters $f = 1$ and $t = 0.001$[11].

H. pylori, *Y. pestis* (clade III). Despite this gene-fusion, the overall operon organization for organisms in clades II and III is identical.

4.3 Gene-expression of metabolic networks

We have analyzed the diauxic shift data from DeRisi *et al.* [5] by PCA/SVD outlined in section 3.3.2 and Fig. fig:svd-a [39]. The genes in the resulting co-expressed gene-vectors after analysis have been categorized in functional classes. Singular-values of mode 2 have been plotted against those of mode 1 and highlighted according to their functional class (Fig. 24). It turned out that the majority of genes are clustered according to their functional classes. For example, almost all genes related to carbohydrate metabolisms (including energy generation) are clustered. Although some exceptional genes could be identified; the β-component of 6-phosphofructokinase (PFK2) and the α-component of fructose-1,6-bisphosphatase (FBP1), which both catalyze reactions between fructose-6-phosphate and fructose-1,6-bisphosphate, but in opposite chemical directions (Fig 25). PFK2 catalyzes the irreversible conversion of fructose-6-phosphate into fructose-1,6-bisphosphate in the glycolysis network for the consumption of glucose and its degradation into pyruvate, which then feeds into the citric acid cycle. FBP1 catalyzes the exact opposite, and also irreversible, reaction changing fructose-1,6-bisphosphate into fructose-6-phosphate. Although not surprising, it is noteworthy that the cellular network, in changing direction of the glycolytic flux during diauxic shift, of course, has to inhibit PFK2 and to activate FBP1 in the same time. In identifying such genetic switches in metabolic networks will further help to understand the modular organization and regulation of cellular networks [9].

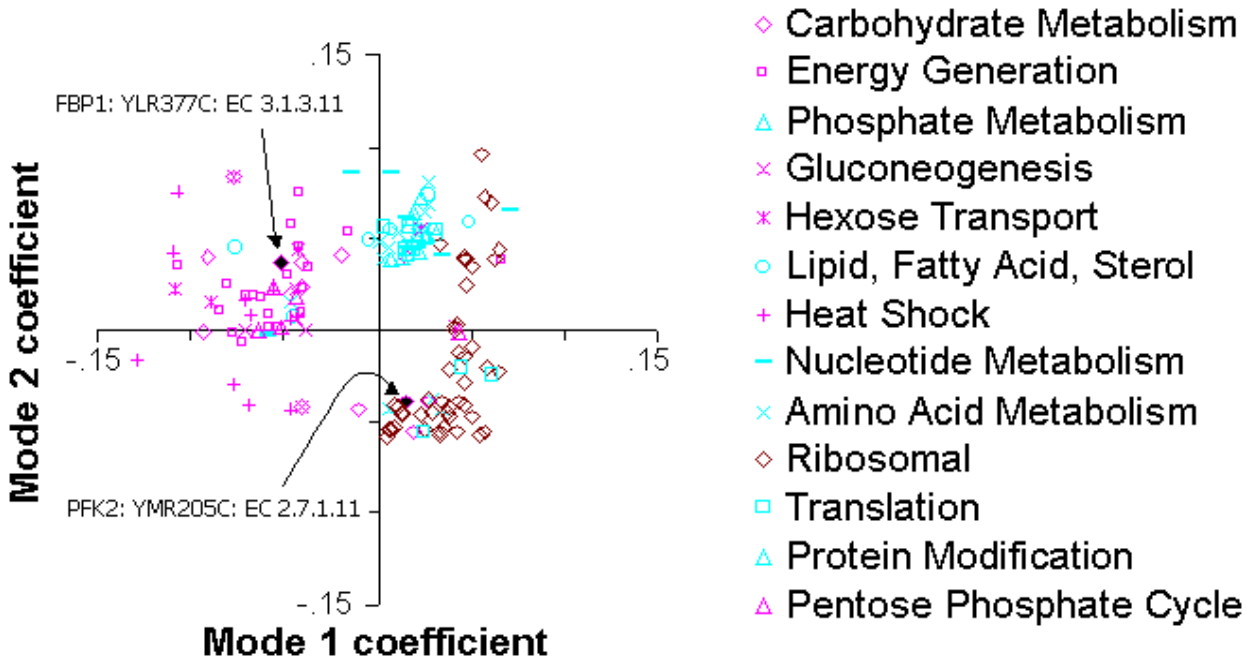

Figure 24: (c.f. [39]) Scatter plot of mode 2 against mode 1. Expressed genes are coded according to their functional class. Similar functional class have identical shades (colors). The two outlined genes in the functional class "carbohydrate metabolism" are counter-expressed (see text)

5 Web-based Information and Databases

A permanently incomplete list of web-based tools and databases relevant to network genomics will be presented.

Enzymes
http://www.brenda.uni-koeln.de (BRENDA)
http://igweb.integratedgenomics.com/EMP (EMP)
http://www.expasy.ch (Expasy)

Gene Expression
http://genex.ncgr.org (GeneX)
http://www.ncbi.nlm.nih.gov/geo (NCBI)
http://genome-www4.stanford.edu/MicroArray/SMD (Stanford Microarray Database)

etabolic Networks
http://ecocyc.PangeaSystems.com (EcoCyc)
http://www.expasy.ch/cgi-bin/search-biochem-index (Expasy: Biochemical Pathways)
http://www.genome.ad.jp/kegg (KEGG)
http://cgsc.biology.yale.edu/metab.html
http://www.gwu.edu/ mpb
http://www.ncgr.org/pathdb (PathDB)
http://wit.mcs.anl.gov/WIT2 (WIT Argonne)
http://wit.integratedgenomics.com/IGwit (WIT/ERGO IntegratedGenomics

eraction Networks
http://dip.doe-mbi.ucla.edu (Database of Interacting Proteins)
http://www.biochem.ucl.ac.uk/bsm/PP/server (Protein-Protein Interaction Server)

gnalling Pathways
http://www.genome.ad.jp/kegg (KEGG)
http://www.grt.kyushu-u.ac.jp/spad (SPAD)

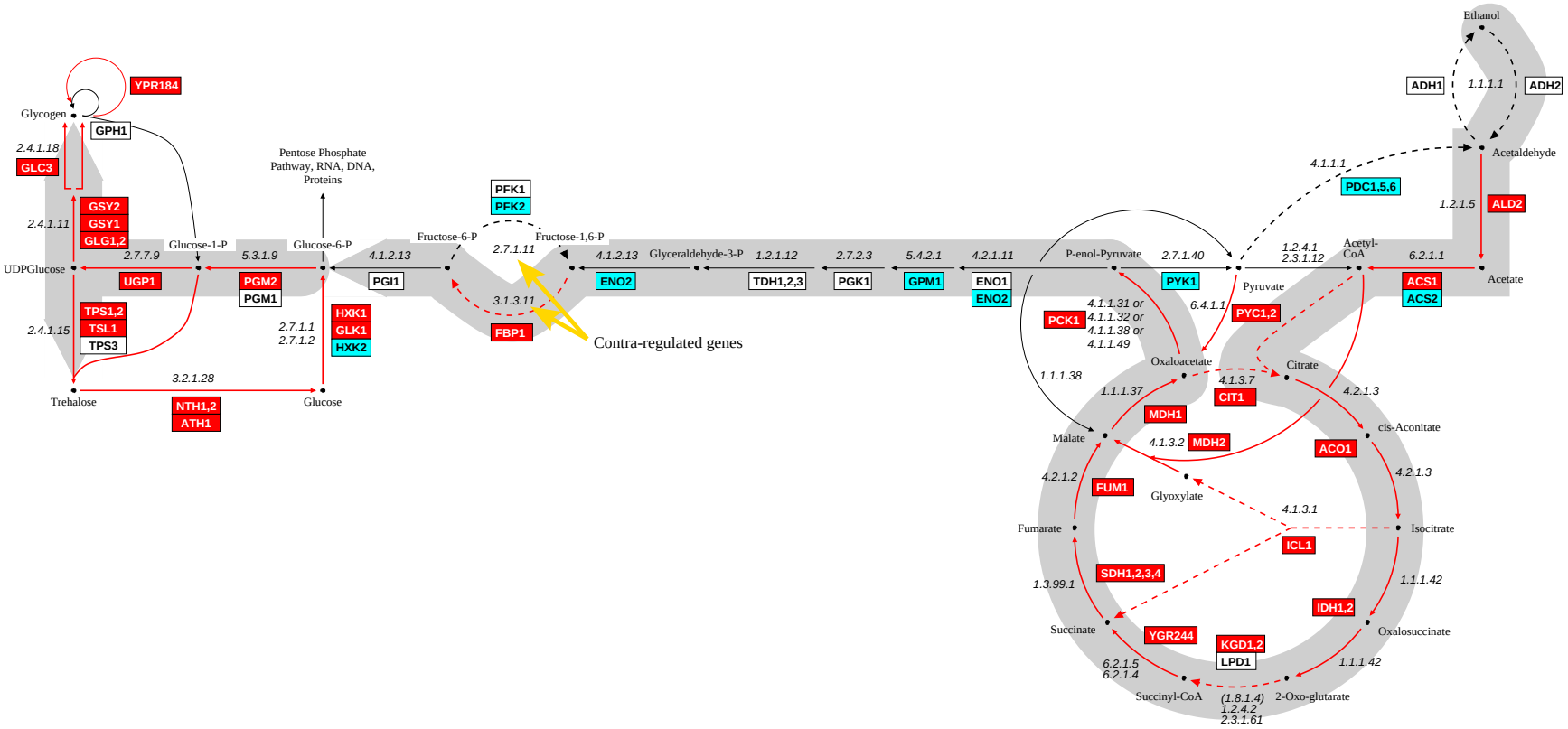

Figure 25: Diauxic shift of yeast. The glycolysis pathway with the citric acid cycle is shown together with gene names and EC number of enzymes functioning in this network. Genes with white gene names on dark (red) background have increased, genes with black names on dark (cyan) background have decreased and genes with black names on white background have unchanged expression after diauxic shift. The heavy grey path indicate the metabolic flux after diaxic shift from ethanol to glucose and starch.

Misc http://www.labmed.umn.edu/umbbd (Biodegradation Database)
http://www-lmmb.ncifcrf.gov/~toms/delila.html (Sequence Logos)

Glossary

CDS CoDing Sequence; a DNA sequene that codes for a protein. In eukaryotic organisms it refers to the spliced mRNA, i.e., it only consists of exons. In the lack of introns, CDS and ORF (see there) are identical.

Cenancestor (c.f. [8]) The most recent common ancestor of the taxa under consideration.

Character (c.f. [8]) Any genic, structural or behavioral feature of an organism having at least two forms of the feature called character states, for example: feather color, red (cardinals) or blue (blue jays); nucleotide, A, T, G or C.

Gene cluster A set of genes withough particular order or direction where each gene is in close distance (typically closer than 300 bp) to neighboring genes (see operon).

Homology (c.f. [8]) The relationship of any two characters that have descended, usually with divergence, from a common ancestral character.

ORF Open Reading Frame; a region on the genome which codes for a protein. In eukaryotic organisms an ORF includes exons and introns (see CDS).

Operon A gene cluster (see there) that is co-expressed by the cellular transcription machinery. Prerequisites for co-expressions and, thus, for an operon are (i) uniform orientation of all genes in operon, (ii) close distance between neighboring genes, and (iii) absence of termination sites.

Orthology (c.f. [8]) The relationship of any two homologous characters whose common ancestor lies in the cenancestor of the taxa from which the two sequences were obtained.

Paralogy (c.f. [8]) The relationship of any two homologous characters arising from a duplication of the gene for that character.

References

[1] O. Alter, P. O. Brown, and D. Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Nat. Acad. Sci. USA*, 2000.

[2] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215:403–410, 1990.

[3] T. Anderson. *Introduction to Multivariate Statistical Analysis.* Wiley & Sons, 2nd edition, 1984.

[4] L. Aravind. Guilt by association: Contextual information in genome analysis. *Genome Res.*, 10:1074–1077, 2000.

[5] J. DeRisi. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278:680–686, 1997.

[6] D. Eisenberg, E. M. Marcotte, I. Xenarios, and T. O. Yeates. Protein function in the post-genomic era. *Nature*, 405:823–826, 2000.

[7] A. J. Enright, I. Iliopoulos, N. C. Kyrpides, and C. A. Ouzounis. Protein interaction maps for complete genomes basede on gene fusion events. *Nature*, 402:86–89, 1999.

[8] W. M. Fitch. Homology: a personal view on some of the problems. *Trends in Genetics*, 16:227–231, 2000.

[9] C. V. Forst. unpublished.

[10] C. V. Forst and K. Schulten. Evolution of metabolism: A new method for the comparison of metabolic pathways using genomic information. *J. Comp. Biol.*, 6:343–360, 1999.

[11] C. V. Forst and K. Schulten. Phylogenetic analysis of metabolic pathways. *J. Mol. Evol.*, 52:471–489, 2001.

[12] G. Golub and C. Van Loan. *Matrix Computations*. John Hopkins Univ. Press, Baltimore, 3rd edition, 1996.

[13] J. Haldane. The origin of life. *Rationalist Ann.*, 148:3–10, 1928.

[14] H. Hartman. Speculations on the origin and evolution. *J. Mol. Evol.*, 4:359–370, 1975.

[15] J. Hughes, P. Estep, S. Tavazoie, and G. church. Computational identification of cis-regulatory elements associated with groups of functionally related genes in saccharomyces cerevisiae. *J. Mol. Biol.*, 296:1205–1214, 2000.

[16] M. Huynen, B. Snel, W. Lathe III, and P. Bork. Predicting protein function by genomic context: Quantitative evaluation and qualitative inferences. *Genome Res.*, 10:1204–1210, 2000.

[17] F. Lipmann. The origin of prebiological systems and of their molecular matrices. pages 259–280. Academic Press, New York, 1965.

[18] J. Liu and C. Lawrence. bayesian inference on bioploymer models. *Bioinformatics*, 15:38–52, 1999.

[19] S. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, San Diego, 2nd edition, 1999.

[20] E. Marcotte, M. Pellegrini, M. Thompson, T. Yeates, and D. Eisenberg. A combined algorithm for genome-wide prediction of protein function. *Nature*, 402:83–86, 1999.

[21] E. M. Marcotte. Computational genetics: finding protein function by nonhomology methods. *Current Opinion Struct. Biol.*, 10:259–365, 2000.

[22] A. M. McGuire and G. M. Church. Predicting regulons and their *cis*-regulatory motifs by comparative genomics. *Nuc. Acids Res.*, 28:4523–4530, 2000.

[23] W. W. Metcalf, J.-K. Zhang, X. Shi, and R. S. Wolfe. Molecular, genetic, and biochemical characterization of the *serc* gene of *methanosarcina barkeri* fusaro. *J. Bact.*, 178(19), 1996.

[24] S. L. Miller. A production of amino acids under possible primitive earth conditions. *Science*, 117:528–529, 1953.

[25] A. Neuwald, J. Liu, and C. Lawrence. Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Sci.*, 4:1618–1632, 1995.

[26] A. I. Oparin. The origin of life. In J. Bernal, editor, *The Origin of Life*. World, Cleveland, 1967. also published in: Proiskhozhdenie Zhizny. IZD Moskovishii Rabochii, Mowcow, 1924.

[27] L. E. Orgel. Evolution of the genetic apparatus. *J. Mol. Biol.*, 38:381–383, 1968.

[28] R. Overbeek, M. Fonstein, M. D'Sousa, G. D. Pusch, and N. Maltsev. The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci. USA*, 96:2896–2901, 1999.

[29] M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proc. Natl. Acad. Sci.*, 96:4285–4288, 1999.

[30] K. R. Popper. *The Poverty of Historicism*. Routledge & Kegan Paul, London, 1957.

[31] K. R. Popper. *Conjectures & Refutations*. Routledge & Kegan Paul, London, 1963.

[32] T. D. Schneider and R. M. Stephens. Sequence logos: A new way to display consensus sequences. *Nucl. Acids Res.*, 18:6097–6100, 1990.

[33] T. F. Smith and M. S. Waterman. Identification of common molecular subspecies. *J. Mol. Biol.*, 147:195–198, 1981.

[34] B. Snel, P. Bork, and M. Huynen. Genome evolution: gene fusion versus gene fission. *Trends Genetics*, 16:9–11, 2000.

[35] G. V. Stauffer. Regulation of serine, glycine, and one-carbon biosynthesis. In K. Herrman and R. L. Sommerville, editors, *Amino Acids: biosynthesis and genetic regulation*, pages 103–113. Addison-Wesley Publishing Co., Reading, Mass., 1983.

[36] R. L. Tatusov, M. Y. Galperin, D. A. Natale, and E. V. Koonin. The cog database: A tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.*, 28:33–36, 2000.

[37] S. Tsoka and C. Ouzounis. Recent developments and future directions in computational genomics. *FEBS Let.*, 480:42–48, 2000.

[38] G. Wächtershäuser. Evolution of the first metabolic cycles. *Proc. Natl. Acad. Sci. USA*, 87:200–204, 1990.

[39] M. Wall, P. Dyck, C. Forst, and T. Brettin. Use of single value decomposition to identify functionally related gene groups from microarray data. *Bioinformatics*, 2001. submitted.

[40] R. Wickner. [ure3] as an altered ure2 protein: evidence for a prion analog in *saccharomyces cerevisiae*. *Science*, 264:566–569, 1994.

[41] Q. Wu and T. Maniatis. A striking organization of a large family of human neural cadherin-like cell adhesion genes. *Cell*, 97:779–790, 1999.

[42] I. Xenarios, D. W. Rice, L. Salwinski, M. K. Baron, E. M. Marcotte, and D. Eisenberg. DIP: the Database of Interacting Proteins. *Nucleic Acids Res.*, 28:1289–1291, 2000.

[43] G. Xie, C. A. Bonner, and R. A. Jensen. A probable mixed-function supraoperon in *pseudomonas* exhibits gene organization features of both inthergenomic conservation and gene shuffling. *J. Mol. Evol.*, 49:108–121, 1999.